

CORGI: Consistency-Aware 3D Dog Reconstruction from a Single Image in the Wild

Yuxiao Wu^{1,†}, Boyi Zhu^{1,†}, Weile Li¹, Yumeng Liu¹(✉), Youcheng Cai¹(✉), XiaoMing Fu¹, and Ligang Liu¹

© The Author(s)

Abstract Reconstructing high-fidelity 3D models of highly articulated animals, such as dogs, from a single in-the-wild image remains a formidable challenge. In this paper, we introduce CORGI, a novel framework for consistency-aware 3D dog reconstruction from a single input image that completely eliminates the need for 3D supervision. To overcome generative inconsistencies and the lack of multi-view capture, our pipeline introduces three core components. First, we propose a Canonical-Driven Orbital Generation (CDOG) strategy, utilizing specialized Canonical and Orbit LoRAs to normalize arbitrary input poses and synthesize reliable pseudo-multi-view images. Second, we design a Consistency-aware Deformable 3DGS (CA-3DGS) module that anchors to D-SMAL: a parametric dog prior, explicitly modeling per-view generative errors through dedicated neural deformation fields to learn accurate vertex-level displacements. Finally, to eliminate structural distortions and recover high-frequency details, we introduce a self-supervised Deformation-Conditioned Generative Repair (DCGR) module. Extensive experiments demonstrate that CORGI achieves state-of-the-art performance, generalizing seamlessly across diverse dog breeds to produce geometrically accurate, visually coherent, and fully animatable 3D assets ready for downstream applications. Project page: <https://dzzzby.github.io/CORGI/>

Keywords 3D Animal Reconstruction, Single-Image Reconstruction, 3D Gaussian Splatting, Cross-View Consistency

1 Introduction

Reconstructing the 3D shape of animals has long been a fundamental challenge in computer vision and computer graphics, supporting a wide range of applications in AR/VR and digital content creation. Among non-human species, dogs constitute a particularly compelling target because, as highly articulated quadrupeds living in close association with humans, they exhibit substantial structural variability and complex motion dynamics that continue to attract significant research interest [1]. Animals are inherently non-cooperative and exhibit continuous non-rigid deformations, making calibrated multi-view capture highly impractical. Consequently, reconstructing high-fidelity 3D representations of dogs in the wild from a single input image with arbitrary pose, viewpoint, and background remains an open and important problem.

Existing methods for 3D animal reconstruction can generally be categorized into template-based and template-free approaches. Template-based approaches employ a parametric 3D template to constrain the ill-posed solution space. For instance, BITE [2] extends the widely adopted SMAL model [3] into the canine domain through the proposed D-SMAL representation, thereby improving pose estimation performance, while AnimalAvatar [4] further enhances dynamic shape tracking from casually captured videos. Nevertheless, constrained by the limited representational capacity of the templates, template-based approaches are inherently limited in capturing subject-specific geometric details. Conversely, template-free approaches [5–7] aim to reconstruct articulated 3D shapes from image collections without relying on predefined templates, thereby offering greater flexibility and applicability. These methods often depend heavily on silhouette or semantic consistency, which often results in overly smooth geometries and insufficient high-frequency details when applied to single in-the-wild image.

Recently, the advent of generative 3D reconstruction has enabled new paradigms for single-view modeling. To overcome the scarcity of multi-view data for animal subjects

¹ School of Mathematical Sciences, University of Science and Technology of China, Hefei, 230026, China. E-mail: W. Wu, wuyx2020@mail.ustc.edu.cn; B. Zhu, dzzzby@mail.ustc.edu.cn; W. Li, liweile@mail.ustc.edu.cn; Y. Liu, lym29@mail.ustc.edu.cn(✉); Y. Cai, caiyoucheng@ustc.edu.cn(✉); X. Fu, fuxm@ustc.edu.cn; L. Liu, lgliu@ustc.edu.cn.

[†] These authors contributed equally to this work and should be considered co-first authors.

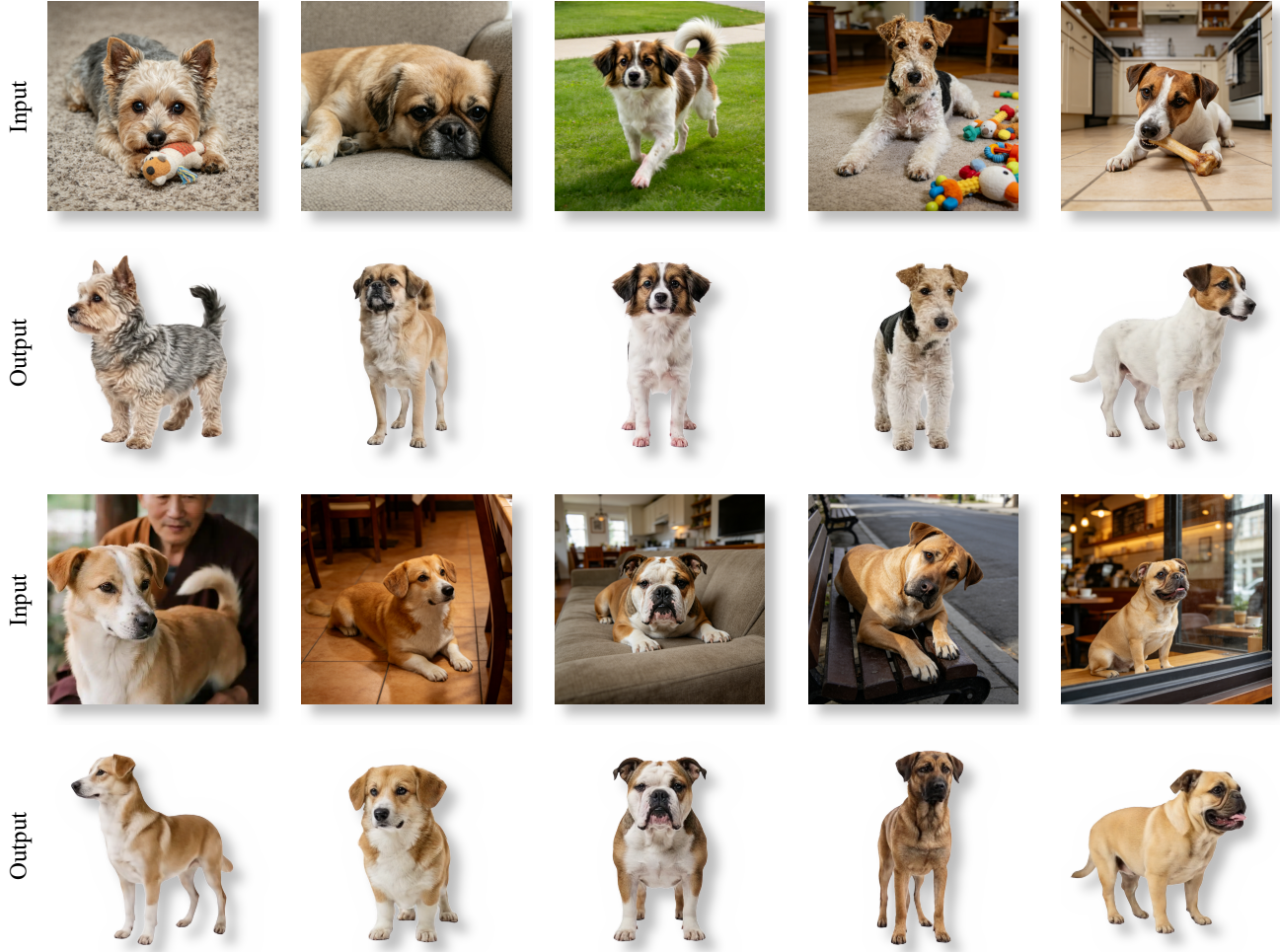


Fig. 1 We propose **CORGI**, a novel framework for reconstructing a high-fidelity 3DGS dog in canonical pose from a single in-the-wild image.

required by traditional reconstruction methods, multi-view diffusion models [8–10] and video diffusion models [11, 12] generate virtual observations from a single image, which are then leveraged by reconstruction pipelines to recover the underlying 3D shape. In the domain of dogs, recent works such as DogRecon [13] leverage the canine prior encoded in D-SMAL to guide multi-view image synthesis, followed by animatable 3D Gaussian Splatting (3DGS) [14] reconstruction. Despite these promising advancements, a critical limitation persists: existing generative models frequently suffer from severe cross-view inconsistencies, leading to blurred textures and geometric artifacts during 3D optimization.

Inspired by the powerful capabilities of modern generative models, we adopt a “generation-then-reconstruction” pipeline to reconstruct high-fidelity 3D dog models from a single in-the-wild image. However, applying this paradigm to highly articulated animals introduces two major challenges. Unlike humans, it is hard to let animals cooperate with scanning

or controlled multi-view capture. As a result, obtaining perfectly aligned real-world multi-view images together with corresponding ground-truth 3D models is extremely difficult, making supervised training impractical. Consequently, we must rely heavily on generative models to synthesize multi-view observations for reconstruction. However, this introduces a second challenge: due to the inherent nature of generative models, the synthesized results often exhibit inconsistent textures and geometries across views, which is undesirable for coherent 3D reconstruction.

To address these challenges, we introduce **CORGI**, a novel framework for consistency-aware 3D dog reconstruction from a single in-the-wild image. Specifically, our pipeline comprises three core components. First, we propose the **Canonical-Driven Orbital Generation (CDOG)** strategy, which transforms a single in-the-wild animal image with an arbitrary pose into a 360-degree orbital sequence anchored at a canonical standing pose for 3D reconstruction,

thereby providing dense and structured multi-view observations for 3D reconstruction. Second, we introduce the **Consistency-Aware Deformable 3DGS (CA-3DGS)** module, which enables robust reconstruction from inconsistent pseudo-multi-view observations by recovering a base 3DGS representation following a dog prior while modeling view-specific inconsistencies with neural deformation fields. Third, because these deformation fields are tied to the observed views and do not generalize well to unseen viewpoints, we propose the **Deformation-Conditioned Generative Repair (DCGR)** module and a novel self-supervised training strategy to repair artifacts in the reconstructed CA-3DGS and improve novel-view synthesis. Together, these designs enable the reconstruction of high-quality animatable 3D dog assets from a single in-the-wild image. Extensive experiments demonstrate that our method achieves state-of-the-art reconstruction performance across the immense structural diversity of dogs; proving that CORGI generalizes seamlessly not just to Corgis, but to diverse breeds like Border Collies, while effectively supporting downstream realistic animation.

In summary, our main contributions are as follows:

- We introduce **CORGI**, a novel system capable of reconstructing high-fidelity 3D dog from a single in-the-wild-image, effectively eliminating the requirement for paired 3D training data.
- Our method addresses the severe ambiguity of monocular 3D reconstruction by using **CDOG** to expand a single image with arbitrary pose into dense observations, and propose **CA-3DGS** to recover a robust initial 3D representation from these generated views despite their cross-view inconsistencies.
- We introduce **DCGR** with a self-supervised learning framework that effectively repairs generative artifacts, producing geometrically accurate and visually coherent 3D models suitable for downstream animation applications.

2 Related work

2.1 Animal 3D Reconstruction

Template-based approaches. Inspired by the monumental success of parametric models in human body digitization, early and foundational efforts in 3D animal reconstruction heavily relied on template-based or parametric priors to constrain the severely ill-posed nature of monocular reconstruction. The pioneering SMAL model [3] established a skinned multi-animal linear formulation, which subsequently catalyzed the development of species-specific adaptations. Notable examples include D-SMAL [2] tailored for the unique

skeletal kinematics of canines, as well as hSMAL [15] and VAREN [16] designed for equines. Building upon these robust geometric priors, a plethora of methods have been proposed to recover articulated shape and pose from single images, sparse multi-view setups, or monocular videos. These frameworks typically employ either computationally intensive optimization-based fitting or efficient feed-forward regression networks [2, 17–19]. More recently, the field has witnessed a paradigm shift toward advanced rendering primitives. For instance, GART [20] elegantly extended the template-based paradigm by replacing traditional mesh surfaces with 3D Gaussian primitives coupled with learnable skinning weights, thereby enabling highly efficient, animatable reconstruction with enhanced rendering quality.

Template-free approaches. Despite their robustness, template-based methods are inherently bottlenecked by the limited topological expressivity of predefined meshes, making them ill-equipped to capture subject-specific geometric variations, such as fluffy fur or distinct ear shapes across different dog breeds. To circumvent these topological constraints, a parallel and flourishing line of research explores template-free reconstruction. These methodologies typically leverage neural implicit representations, articulated neural parts, or canonical-to-posed feature decompositions to learn category-level priors directly from data. For instance, BANMo [21] pioneered the disentanglement of canonical shape and non-rigid motion to learn animatable neural implicit models from monocular videos. Scaling this concept up, MagicPony [7] and 3D-Fauna [5] successfully extracted category-level articulated models directly from in-the-wild image collections. Concurrently, part-based frameworks—such as LASSIE [6], Hi-LASSIE [22], LEAPARD [23], and ARTIC3D [24]—approach the problem by discovering semantic correspondences and assembling animals through self-supervised articulated neural parts. Further broadening the design space, methods like CASA [25] and DualPM [26] introduced category-agnostic skeletal reasoning and canonical point-map representations to handle extreme topological diversity.

Despite this rapid progress, reconstructing highly articulated animals from in-the-wild images remains profoundly challenging. The scarcity of large-scale, high-quality 3D ground truth paired with diverse in-the-wild images forces existing methods to rely on synthetic data, toy scans, or weak 2D supervision. Consequently, both template-based and template-free approaches often struggle to generalize across the vast structural variations of different dog breeds, suffering severe geometric degradation when confronted with complex

poses, self-occlusions, and fine-scale details like fur.

2.2 Single-Image 3D Generation

Recovering 3D content from a single image has long been a fundamental, albeit inherently ill-posed, problem in computer vision and graphics [27]. Early approaches typically formulated this as a supervised reconstruction task, training models to directly regress 3D geometry, shape parameters, or category-specific representations from a single input image [28–32]. While these pioneering methods established the foundational paradigm of single-image 3D reconstruction, they were often bottlenecked by their reliance on restrictive object categories and limited geometric representations. Beyond direct reconstruction, a crucial parallel line of work introduced 3D-aware neural rendering and generative representations, which subsequently became vital priors for image-conditioned 3D generation. Methods such as GRAF [33], GIRAFFE [34], StyleNeRF [35], and EG3D [36] demonstrated that radiance-field-based or geometry-aware representations could synthesize view-consistent images while effectively capturing the underlying 3D structure.

With the proliferation of diffusion models, a highly influential line of research emerged that leverages pretrained 2D diffusion models to optimize an underlying 3D representation via diffusion guidance or score distillation. DreamFusion [37] introduced Score Distillation Sampling (SDS) to optimize a neural radiance field using a 2D diffusion prior. Follow-up frameworks, including LucidDreamer [38], Magic123 [39], and ProlificDreamer [40], further refined this approach by improving geometry initialization, optimization stability, and overall visual fidelity. Although these optimization-based methods can produce exceptionally high-quality results, they inherently require computationally expensive per-instance optimization, rendering them less suitable for scalable, feed-forward reconstruction.

To overcome the efficiency bottleneck of per-instance optimization, another major direction reformulates single-image 3D generation as a multi-view synthesis problem. Zero-1-to-3 [8] demonstrated that diffusion models could synthesize plausible novel views from a single input image, catalyzing a new pipeline that first generates sparse views and subsequently recovers 3D geometry. Later methods such as SyncDreamer [9], MVDream [41], and Wonder3D [42] significantly improved cross-view consistency and geometric coherence in the hallucinated views. Building upon this multi-view paradigm, recent Large Reconstruction Models (LRMs) have substantially advanced inference efficiency by amortizing the reconstruction process into a single feed-forward network,

including LRM [43], PF-LRM [44], CRM [45], LGM [46], InstantMesh [47], TripoSR [48], and One-2-3-45 [49].

Beyond optimization-based techniques and multi-view diffusion pipelines, the most recent frontier explores native 3D generative models aimed at modeling 3D assets directly within compact latent spaces. Representative examples include CLAY [50], 3DShape2VecSet [51], TRELLIS/TRELLIS-2 [52], and Hunyuan3D [53]. These architectures learn structured latent spaces over native 3D representations and perform image-conditioned generation directly within those spaces. Compared to earlier pipelines, these approaches offer superior scalability and exhibit an increasingly generalized capability for 3D generation.

Despite this sweeping and rapid progress across multiple paradigms, reconstructing highly articulated animals such as dogs from a single input image remains particularly formidable. A central bottleneck is the conspicuous absence of large-scale paired supervision between real animal images and high-quality 3D assets. In practice, much of the available 3D animal data used to train modern image-to-3D models originates from synthetic repositories, static game assets, or animation models, rather than genuine 3D captures aligned with real-world images. Consequently, even the most robust general-purpose single-image-to-3D models encounter a severe domain gap when applied to in-the-wild dog images. This discrepancy inevitably leads to drastically reduced reconstruction fidelity, poor articulation accuracy, and a loss of visual realism. By explicitly addressing these domain gaps and structural inconsistencies, our proposed framework aims to bridge the divide between generative priors and highly articulated real-world subjects.

3 Preliminaries

3.1 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [14] has recently emerged as an efficient explicit point-based representation for novel view synthesis and 3D scene reconstruction. 3DGS models a continuous 3D scene using a collection of unstructured, anisotropic 3D Gaussians.

Each 3D Gaussian is characterized by a center position μ , a 3D covariance matrix Σ , an opacity α , and view-dependent color features $\mathbf{c}$, which are typically encoded using spherical harmonics (SH). The spatial contribution of a 3D Gaussian at a point $\mathbf{x}$ is defined as:

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right). \quad (1)$$

The covariance matrix Σ is decomposed into a rotation matrix $\mathbf{R}$ and a scaling matrix $\mathbf{S}$, which are parameterized by

a unit quaternion and a 3D scaling vector, respectively:

$$\Sigma = \mathbf{R} \mathbf{S} \mathbf{S}^T \mathbf{R}^T. \quad (2)$$

For efficient rendering, 3DGS employs an optimized tile-based rasterizer. Given a camera view transformation $\mathbf{W}$ and the Jacobian matrix $\mathbf{J}$ corresponding to the affine approximation of the projective transformation, the projected 2D covariance matrix Σ_{2D} is computed as follows [54]:

$$\Sigma_{2D} = \mathbf{J} \mathbf{W} \Sigma \mathbf{W}^T \mathbf{J}^T. \quad (3)$$

The final color $\mathbf{C}$ of a pixel is obtained by sorting the projected Gaussians in front-to-back order and performing point-based α -blending, thereby approximating the volume-rendering integral:

$$\mathbf{C} = \sum_{i=1}^N c_i \alpha'_i \prod_{j=1}^{i-1} (1 - \alpha'_j), \quad (4)$$

where c_i denotes the color of the i -th Gaussian encoded via spherical harmonics, α'_i represents the effective 2D opacity obtained by multiplying the learned opacity α_i with the corresponding 2D Gaussian value at the pixel location.

3.2 Diffusion-Guided Repair

While 3D Gaussian Splatting provides efficient and photorealistic novel-view synthesis for viewpoints close to the original camera trajectories, it inherently struggles with under-constrained regions and sparse-view settings, often producing severe artifacts such as spurious geometry, floaters, and holes [55, 56]. To alleviate these limitations, recent approaches have increasingly exploited the strong generative priors of pre-trained 2D diffusion models to repair artifacts and hallucinate plausible details in unobserved regions.

The repair process is commonly formulated as a conditional image synthesis task. Given a degraded image I_{render} rendered from an imperfect 3D representation, the diffusion model aims to model the conditional distribution $p(I_{gt}|I_{render})$ in order to generate a high-fidelity, artifact-free image $\hat{I}$. Mathematically, a conditional denoising network ϵ_θ is trained to predict the noise ϵ added to the latent representation z_t . At timestep t , the network predicts the injected noise ϵ from the noisy latent representation z_t by optimizing the standard denoising score-matching objective:

$$\mathcal{L}_{DM} = \mathbb{E}_{z_0, \epsilon \sim \mathcal{N}(0, I), t} [\|\epsilon - \epsilon_\theta(z_t, I_{render}, t)\|_2^2]. \quad (5)$$

4 Method

As illustrated in Figure 2, we propose **CORGI**, a “generation-then-reconstruction” framework designed to reconstruct high-fidelity, animatable 3D dog models from a single in-the-wild

input image without relying on any 3D supervision. To overcome the inherent ill-posedness and cross-view inconsistencies of single-view generation, our pipeline seamlessly integrates three core components. First, the **Canonical-Driven Orbital Generation (CDOG)** strategy employs specialized Canonical and Orbit LoRAs to normalize the arbitrary input pose and synthesize a reliable 360-degree pseudo-multi-view images. Subsequently, these 2D observations are lifted into 3D via the **Consistency-aware Deformable 3DGS (CA-3DGS)** module, which anchors 3D Gaussian primitives to a D-SMAL template and optimizes vertex-level displacements, while explicitly absorbing view-dependent generative errors through dedicated neural deformation fields. Finally, to rectify residual geometric distortions and recover high-frequency textures, we introduce the **Deformation-Conditioned Generative Repair (DCGR)** module, which leverages a 2D continuous deformation map as geometric conditioning for a pre-trained diffusion model to refine the overall 3D representation under a novel self-supervised training strategy.

4.1 Canonical-Driven Orbital Generation

Acquiring perfectly aligned, multi-view captures of dynamic animals in the wild is notoriously difficult, making direct 3D supervision virtually impossible. To overcome this, we leverage an image generation model to transform an input image from an arbitrary pose and cluttered background into a canonical pose with a clean background, and subsequently employ a video generation model to synthesize a 360-degree orbit video for 3D reconstruction. A straightforward strategy to circumvent this data scarcity would be to rely heavily on test-time prompt engineering, manually adjusting text prompts conditioned on the input image to coax a pre-trained generative model into synthesizing the desired outputs. However, such heuristic prompt tuning is highly inefficient and notoriously unstable. To address this problem, we shift the paradigm from unpredictable test-time prompting to systematically fine-tuning pre-trained generative models with Low-Rank Adaptation (LoRA) [57]. Our Canonical-Driven Orbital Generation (CDOG) module decomposes the inherently ill-posed single-view reconstruction problem into two tractable, stage-wise generative processes: Canonical Pose Generation and Orbital Video Generation.

4.1.1 Canonical Pose Generation

The objective of this stage is to transform an input image, which captures a dog in an arbitrary pose and cluttered environment, into a canonical pose image, characterized by a canonical standing pose and a clean background.

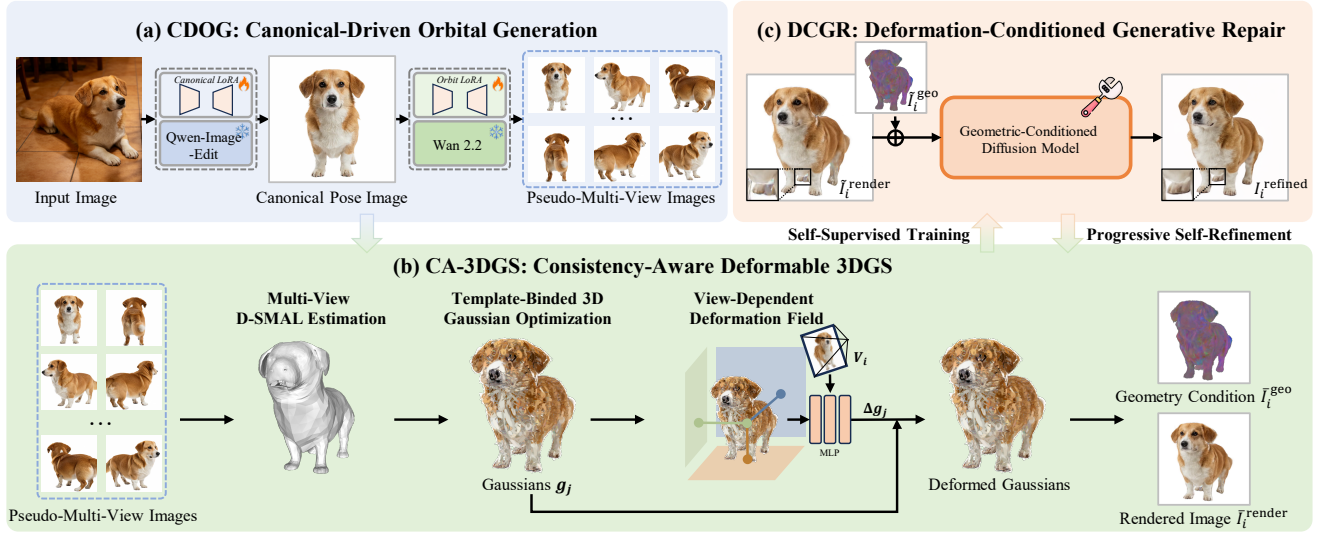


Fig. 2 Overview of the CORGI framework. From a single input image, CORGI reconstructs a high-fidelity, animatable 3D dog without 3D supervision. (a) **CDOG** normalizes the input pose and synthesizes a reliable 360-degree pseudo-multi-view images. (b) **CA-3DGS** lifts these 2D observations into a deformable 3DGS field anchored to a D-SMAL template, using neural deformation fields to explicitly isolate view-dependent generative errors. (c) **DCGR** leverages rendered deformation maps to geometrically condition a diffusion model, rectifying residual artifacts and recovering high-frequency details in a self-supervised manner.

Dataset. To enable consistent pose transformation, we introduce a self-supervised data curation paradigm. Utilizing the advanced generative capabilities of Qwen-Image-Edit [58] guided by meticulously engineered prompts, we efficiently generate and filter a curated dataset containing 100 high-quality arbitrary-canonical pose image pairs. Each pair consists of a canine subject in a natural, arbitrary pose alongside its exact identity counterpart in a canonical standing pose with a clean background.

Framework. Building upon the robust zero-shot editing capabilities of Qwen-Image-Edit [58] as our baseline, we formulate the pose normalization as a conditional image-to-image translation task. Using our self-supervised paired dataset, we fine-tune a *Canonical LoRA* within the transformer backbone of the baseline model. This explicit fine-tuning forces the network to learn the conditional mapping from the arbitrary input to the canonical pose, effectively disentangling the subject’s intrinsic identity and high-frequency textures from its original arbitrary pose. This provides a reliable spatial anchor for the subsequent orbital video generation.

4.1.2 Orbital Video Generation

Given the canonical pose image, this stage synthesizes a dense sequence of multi-view observations to mimic a 360-degree camera flythrough.

Dataset. Due to the scarcity of high-quality orbital videos specifically featuring dogs, we deliberately assemble a specialized training dataset comprising 100 high-quality orbital

videos that include a mixture of canine and *non-canine* subjects. Notably, our primary objective is for the model to exclusively learn the pure geometric prior of a 360-degree azimuthal camera trajectory. Therefore, the inclusion of non-canine data provides excellent supervisory signals for camera motion without being strictly tied to the subject’s semantics.

Framework. We inject an *Orbit LoRA* into a pre-trained Image-to-Video (I2V) diffusion backbone, Wan 2.2 [59], and train it exclusively on our curated orbital dataset. This training strategy successfully decouples the spatial-temporal camera motion from the semantic identity of the subject. Finally, to enforce the cyclic consistency required by the subsequent 3DGS optimization, we utilize a dual-frame conditioning mechanism, setting both the initial and terminal frames strictly to the generated canonical pose image. This forces the diffusion trajectory to synthesize a perfect closed-loop orbital video sequence, providing robust and consistent pseudo-multi-view images.

4.2 Consistency-Aware Deformable 3DGS

Given the pseudo-multi-view images $\{I_i^{gt}\}_{i=1}^N$ generated by the CDOG module, our goal is to reconstruct a high-fidelity canonical 3DGS representation of the subject. While standard 3DGS demonstrates remarkable rendering quality and efficiency for static scenes, directly applying it to generatively synthesized multi-view data is inherently problematic. Diffusion-generated videos often exhibit view-dependent inconsistencies, texture flickering, and local geometric shifts. To

address these issues, we propose the Consistency-aware Deformable 3DGS (CA-3DGS) representation. CA-3DGS binds the 3D Gaussian primitives to a parametric canine template (D-SMAL) and explicitly models generative inconsistencies through a view-dependent neural deformation field.

4.2.1 Multi-View D-SMAL Estimation

Unlike prior single-view methods [2, 13], we utilize the synthesized dense multi-view sequence for accurate template alignment. We estimate camera parameters $\{V_i\}_{i=1}^N$ using COLMAP and initialize the D-SMAL template with BITE [2]. We denote $\mathcal{R}_i(\cdot)$ as the differentiable renderer associated with the camera parameters of the i -th view. We represent the D-SMAL parameters as $S = (\beta, \theta, \gamma, \phi, \rho)$, where β denotes the shape parameters, θ denotes the 6D rotations of all the joints, γ denotes the global translation, ϕ denotes the global orientation, and ρ is a global scaling factor to accommodate dogs of varying sizes. This parameterization follows DogMo [60]. We denote the set of all vertices of the resulting D-SMAL template by X and the 3D articulated joints by J . We optimize the parameters jointly across all N views by minimizing the combined objective $\mathcal{L}_{\text{D-SMAL}} = \mathcal{L}_{mk} + \mathcal{L}_{kp}$. The two terms are defined as follows.

Soft Mask Loss. The term $\mathcal{L}_{mk}$ denotes the soft mask loss, which encourages the projected mesh to align with the multi-view silhouettes:

$$\mathcal{L}_{mk} = \sum_{i=1}^N \|M_i - \mathcal{R}_i^{\text{soft}}(X)\|_2,$$

where M_i denotes the foreground pixel set of the segmentation mask in the i -th view, and $\mathcal{R}_i^{\text{soft}}(X)$ denotes the soft mask rendering of the mesh vertices X in the i -th view as introduced in [61]. Unlike a binary silhouette, the soft mask outputs a continuous probability value per pixel, indicating the likelihood of being covered by the projected mesh. This differentiability allows gradient flow through the mask loss.

Keypoint Loss. The term $\mathcal{L}_{kp}$ denotes the sparse keypoint loss, which constrains the articulated pose through sparse keypoint supervision:

$$\mathcal{L}_{kp} = \sum_{i=1}^N \|p_i - \mathcal{P}_i(J)\|_2, \quad (6)$$

where p_i denotes the 2D keypoints detected by BARC [18], and $\mathcal{P}_i(J)$ represents the projection of the corresponding 3D joints.

4.2.2 Template-Binded 3D Gaussian Optimization

To maintain topological consistency and facilitate downstream animation, we explicitly attach the 3D Gaussian primitives to the surface of the optimized D-SMAL mesh. Inspired by

[62, 63], we sample Gaussians uniformly over the template surface within its 2D UV parameterization space. Let μ^{2D} denote the UV coordinates of a Gaussian. Its base 3D position $U(X, \mu^{2D})$ is computed through barycentric interpolation of the mesh vertices X . To capture subject-specific non-rigid details (e.g., fur) that cannot be represented by the base template, we introduce a learnable scalar τ that displaces the Gaussian along the interpolated surface normal $\mathbf{n}$. The final position of the j -th Gaussian is formulated as:

$$\mu_j = U(X, \mu_j^{2D}) + \tau_j \mathbf{n}_j. \quad (7)$$

Crucially, in addition to the Gaussian attributes, the underlying 3D vertices of the D-SMAL mesh are jointly optimized as learnable parameters to better fit subject-specific geometry.

4.2.3 View-Dependent Deformation Field

Although the orbital sequence provides dense 360-degree coverage, the generated pseudo-multi-view images exhibit local cross-view inconsistencies. Forcing a static 3DGS model to fit these inconsistent observations leads to severe texture blurring and floating artifacts. To disentangle the underlying canonical geometry from the view-dependent inconsistencies introduced by the generative model, we represent the scene using a base 3DGS model $\mathcal{G}$ together with a view-dependent neural deformation field $\mathcal{F}_\theta$. In this way, $\mathcal{G}$ captures the shared, view-consistent geometry, while $\mathcal{F}_\theta$ accounts for the view-dependent variations in the Gaussian attributes induced by the inconsistent observations.

Here we denote the base model as $\mathcal{G} = \{g_j\}_{j=1}^K$, where each Gaussian is parameterized by $g_j = (\mu_j, s_j, r_j, \alpha_j, c_j)$. Specifically, μ_j denotes the position defined in Eq. (7), s_j and r_j represent the scale and rotation of the j -th Gaussian respectively, α_j and c_j are the opacity and color mentioned in Eq. (4).

The deformation field $\mathcal{F}_\theta$ is parameterized by an efficient triplane representation [36] for spatial feature encoding and a lightweight Multi-Layer Perceptron (MLP) for residual deformation decoding. Given a camera view V_i , the deformation field $\mathcal{F}_\theta$ takes the position μ_j of the j -th Gaussian as input and predicts a residual offset for its attributes:

$$\Delta g_{j,i} = \mathcal{F}_\theta(V_i, \mu_j). \quad (8)$$

The deformed attributes used for rendering the i -th view are updated as:

$$g'_{j,i} = g_j + \Delta g_{j,i}. \quad (9)$$

This design effectively represents the inconsistencies by the neural deformation field while preserving a clean 3D representation shared across all views.

For i -th view, the image is rendered from the deformed attributes $\{g'_{j,i}\}_{j=1}^K$ using the standard differentiable 3DGS rasterizer. We jointly optimize the mesh vertices X , the Gaussian attributes $\{\tau_j, s_j, r_j, \alpha_j, c_j\}_{j=1}^K$, and parameters of the deformation field F_θ using the following objective:

$$\mathcal{L} = \mathcal{L}_{\text{image}} + \lambda_{\text{arap}} \mathcal{L}_{\text{arap}} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}, \quad (10)$$

where each loss is defined as follow. Firstly, $\mathcal{L}_{\text{image}}$ consists of L_1 loss and Structural Similarity Index Measure (SSIM) to enforce pixel-wise and perceptual consistency with respect to $\{I_i^{\text{gt}}\}_{i=1}^N$. Secondly, $\mathcal{L}_{\text{arap}}$ denotes As-Rigid-As-Possible (ARAP) energy [64], which encourages local rigidity by constraining each vertex neighborhood to preserve its original shape up to a local rotation. Specifically, it is computed by penalizing discrepancies between the deformed edge vectors and the corresponding template edge vectors after optimal local rotation alignment. Finally, $\mathcal{L}_{\text{reg}}$ regularizes the deformation field to encourage sparse and smooth corrections:

$$\mathcal{L}_{\text{reg}} = \frac{1}{NK} \sum_i \sum_j^K \|\Delta g_{j,i}\|_2^2 + \lambda_{\text{grad}} \mathcal{L}_{\text{grad}}, \quad (11)$$

where K is the total number of Gaussian primitives. The first term penalizes the magnitude of the deformations to suppress large and unstable deformations. The second term, $\mathcal{L}_{\text{grad}}$, penalizes spatial variations in the deformation residuals, whose gradients are approximated using finite differences on the triplane grid, thereby explicitly enforcing spatial smoothness.

4.3 Deformation-Conditioned Generative Repair

Although the view-dependent deformation field $\mathcal{F}_\theta$ absorbs per-view inconsistencies during optimization, its predictions are inherently tied to the input view. If we render each view V_i using its corresponding deformation, the resulting 3DGS would reproduce the view inconsistencies present in the generated inputs. Therefore, instead of applying view-specific deformations at inference time, we freeze the neural deformation field to a single reference view (typically the first view V_0) and use the resulting deformed geometry for rendering all novel views. Ideally, multi-view consistent images $\{\tilde{I}_i^{\text{render}}\}_{i=1}^N$ should be rendered by the differentiable renderer $\mathcal{R}_i$, where

$$\tilde{I}_i^{\text{render}} = \mathcal{R}_i\left(\{g_j + \mathcal{F}_\theta(V_0, \mu_j)\}_{j=1}^K\right). \quad (12)$$

However, because the deformation field $\mathcal{F}_\theta(V_0, \cdot)$ is inherently designed to fit the specific generative flaws of the reference view V_0 , forcing it to adapt to a drastically different view V_i inevitably causes severe degradation, such as geometric tearing and texture blurring. To mitigate these artifacts, we introduce the Deformation-Conditioned Generative Repair

(DCGR) module, which leverages a custom self-supervised learning paradigm to restore degraded image to high-fidelity, artifact-free results.

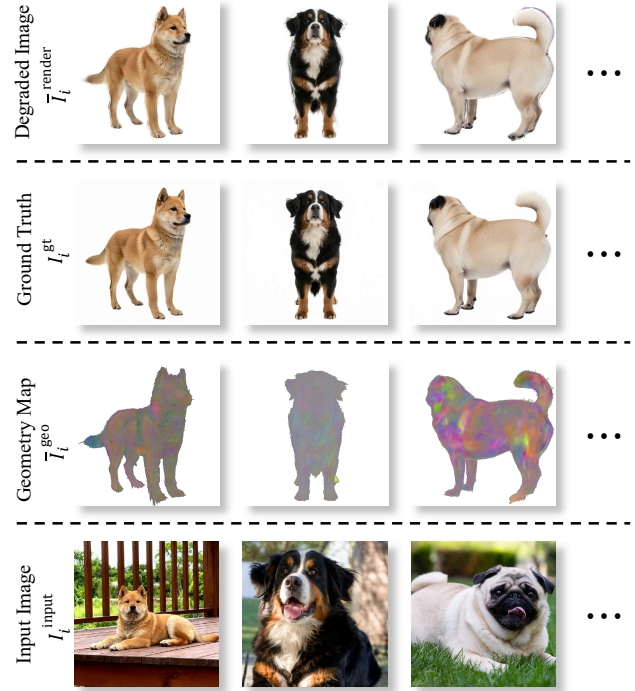


Fig. 3 Creation of Training Dataset. From top to bottom, each row respectively represents: the degraded image $\tilde{I}_i^{\text{render}}$, the ground truth I_i^{gt} , the geometry condition $\tilde{I}_i^{\text{geo}}$, and the input image I_i^{input} as an appearance reference.

4.3.1 Self-Supervised Dataset Creation

To train the DCGR module for correcting generative artifacts under the perspective-locked rendering formulation in Eq. 12, we construct a self-supervised dataset that captures the characteristic degradations produced by the initial 3DGS model without requiring clean 3D ground truth.

To ensure sufficient diversity of real images in the training dataset, we collect a large number of in-the-wild dog images from the Internet, denoted as I^{input} . We then apply the CDOG module to generate pseudo multi-view images $\{\tilde{I}_i^{\text{gt}}\}_{i=1}^N$, which serve as pseudo ground truths, and optimize the CA-3DGS model along with its view-dependent deformation field $\mathcal{F}_\theta$. Hereafter, $\{g_j\}_{j=1}^K$ denotes the attributes of the optimized Gaussians. Because the CA-3DGS model and the deformation field $\mathcal{F}_\theta$ are optimized from inconsistent multi-view images, the resulting renderings naturally contain artifacts. For each camera view V_i , we render the degraded image $\tilde{I}_i^{\text{render}}$ by using V_i as the input to $\mathcal{F}_\theta$:

$$\tilde{I}_i^{\text{render}} = \mathcal{R}_i\left(\{g_j + \mathcal{F}_\theta(V_i, \mu_j)\}_{j=1}^K\right). \quad (13)$$

Notably, $\{\tilde{I}_i^{\text{render}}\}_{i=1}^N$ are multi-view inconsistent renderings

that serve as the degraded images for DCGR, while the corresponding pseudo ground truths $\{I_i^{\text{gt}}\}_{i=1}^N$ are used as supervision targets.

Meanwhile, to provide explicit structural guidance regarding these geometric artifacts, we project the view-specific displacement vectors onto the image plane to obtain a geometry map:

$$\bar{I}_i^{\text{geo}} = \mathcal{R}_i\left(\{\mu'_{j,i}, s'_{j,i}, r'_{j,i}, \alpha'_{j,i}, \Delta\mu_{j,i}\}_{j=1}^K\right), \quad (14)$$

where $\mu'_{j,i}$, $s'_{j,i}$, $r'_{j,i}$, and $\alpha'_{j,i}$ are defined in Eq. (9).

Consequently, for each view i , we construct a self-supervised training quadruplet $\{\bar{I}_i^{\text{render}}, I_i^{\text{gt}}, \bar{I}_i^{\text{geo}}, I_i^{\text{input}}\}$, where $\bar{I}_i^{\text{render}}$ is the degraded image, I_i^{gt} is the pseudo ground truth, the geometry map $\bar{I}_i^{\text{geo}}$ serves as the geometric condition, and the input image I_i^{input} provides the appearance reference.

4.3.2 Network Framework and Progressive Self-Refinement

Geometric-Conditioned Diffusion Model. We build our repair framework upon the single-step diffusion architecture of DiFix3D+ [56], formulating the artifact removal as a geometrically-conditioned translation task. To adapt the generic diffusion prior to our canine reconstruction pipeline, we freeze the base weights and inject a lightweight LoRA module into the backbone.

(1) *Training.* During training, only the inserted LoRA parameters are optimized while the pretrained DiFix3D+ backbone remains frozen. Given a self-supervised training quadruplet $\{\bar{I}_i^{\text{render}}, I_i^{\text{gt}}, \bar{I}_i^{\text{geo}}, I_i^{\text{input}}\}$, the degraded image and the geometry map are concatenated as the structural condition, whereas the input image provides the appearance guidance. The diffusion model is trained to reconstruct the corresponding pseudo ground truth I_i^{gt} using the standard single-step denoising objective adopted in DiFix3D+:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_t, \epsilon} [\|\epsilon - \epsilon_\theta(z_t, c)\|_2^2], \quad (15)$$

where z_t denotes the noisy latent at diffusion step t , ϵ is the sampled Gaussian noise, and

$$c = \{\bar{I}_i^{\text{render}}, \bar{I}_i^{\text{geo}}, I_i^{\text{input}}\} \quad (16)$$

represents the geometric and appearance conditions. Since only the LoRA layers are optimized, the pretrained diffusion prior is efficiently adapted to geometry-aware artifact removal while preserving its powerful image restoration capability.

(2) *Inference.* During inference, given a rendered image $\bar{I}_i^{\text{render}}$ from the current CA-3DGS model, we first generate its corresponding geometry map $\bar{I}_i^{\text{geo}}$ by projecting the view-dependent Gaussian displacements onto the image plane. The rendered image, geometry map, and appearance reference

I_i^{input} are then jointly fed into the LoRA-adapted diffusion model to directly predict a repaired image $\tilde{I}_i^{\text{refined}}$ through a single denoising step. Benefiting from the complementary geometric and appearance guidance, the repaired result preserves the identity and fine-grained texture of the input dog while effectively eliminating rendering artifacts.

Progressive Self-Refinement. To propagate these 2D image enhancements back into the 3D space, we wrap the LoRA-adapted DiFix3D+ network within an progressive self-refinement loop. Starting from the initial representation $\mathcal{G}^0$, which is the optimized CA-3DGS obtained in Section 4.2, each refinement round t executes three sequential operations:

(1) *Rendering.* We render the multi-view images $\{\tilde{I}_{i,t}^{\text{render}}\}_{i=1}^N$ from the current model $\mathcal{G}^t$ according to Eq. 12, while fixing the deformation field to the reference view V_0 to enforce global consistency.

(2) *Generative Repair.* For each camera view V_i , the rendered image $\tilde{I}_{i,t}^{\text{render}}$, its geometry map $\tilde{I}_{i,t}^{\text{geo}}$, and the appearance reference I_i^{input} are jointly fed into our geometric-conditioned diffusion model to produce a refined, artifact-free observation $I_{i,t}^{\text{refined}}$.

(3) *Self-Refinement.* The original pseudo-multi-view images generated by CDOG are replaced with these repaired images to serve as new pseudo ground truths, and the 3DGS parameters are re-optimized using the losses defined in Section 4.2 to yield the updated model $\mathcal{G}^{t+1}$. In practice, we find that performing the self-refinement loop up to $t = 3$ is sufficient to obtain a high-fidelity and fully consistent 3DGS model of the dog.

5 Experiments

5.1 Experiment Setup

Implementation details. The proposed CORGI framework is implemented in PyTorch. For the Canonical-Driven Orbital Generation (CDOG) module, we generate an 80-frame surrounding video to provide dense pseudo-multi-view images for the subsequent reconstruction stage. The Consistency-Aware Deformable 3DGS (CA-3DGS) module is optimized for a total of 30,000 iterations. The regularization weights for the deformation field are empirically set to $\lambda_{\text{arap}} = 0.01$, $\lambda_{\text{reg}} = 0.01$, and $\lambda_{\text{grad}} = 1$. For the Deformation-Conditioned Generative Repair (DCGR) module, the progressive self-refinement loop is executed for 3 rounds, with each round undergoing 8,000 iterations of re-optimization. All experiments are conducted on a single NVIDIA RTX 4090 (48GB) GPU.

Datasets. To rigorously evaluate our framework across both constrained and unconstrained scenarios, we curate two distinct evaluation datasets: (1) **Dog Synthetic Dataset:** This

dataset comprises 10 high-quality, fully rigged 3D canine models. For each model, we render 24 views along a 360-degree orbital trajectory. The first frame strictly serves as the monocular input, while the remaining 23 views are reserved exclusively as ground truth (GT) to quantitatively evaluate multi-view rendering accuracy and geometric fidelity. (2) **Dog Wild Dataset:** To assess in-the-wild generalization capability, we collect 150 real-world dog images sourced from the internet. This dataset covers a vast spectrum of dog breeds, arbitrary poses, varying illuminations, and cluttered backgrounds. Since no 3D ground truth is available for real images, this dataset is used exclusively to evaluate single-view 3D generation quality and robustness.

Metrics. Due to the generative nature of our pipeline, we utilize both reference-based and non-reference metrics to provide a comprehensive and human-perception-aligned evaluation: (1) **Reference-based Metrics:** Evaluated on the Dog Synthetic Dataset, we measure multi-view consistency and reconstruction accuracy against the ground truth using PSNR, SSIM, and LPIPS [65]. Specifically, PSNR evaluates the absolute pixel-level reconstruction error. SSIM quantifies the structural and luminance fidelity of the generated geometries. LPIPS assesses deep feature-level perceptual similarity, which aligns more closely with human visual judgment than traditional pixel-wise metrics. (2) **Non-reference Metrics:** Evaluated on the Dog Wild Dataset, we employ widely adopted perceptual metrics including FID [66] and NIQE [67] to assess rendering realism without ground truth. Specifically, FID quantifies the distribution distance between the rendered images and the real data domain, reflecting overall generative fidelity. NIQE measures the naturalness of the synthesized views based on statistical deviations from natural scene models. Furthermore, because preserving the original subject’s identity is crucial, we follow prior work [68] and calculate the CLIP-cosine similarity between the input image and the rendered novel views to rigorously evaluate semantic and identity consistency.

5.2 Baselines

To demonstrate the superiority of CORGI, we compare our framework against a comprehensive suite of state-of-the-art baselines. These baselines encompass a wide spectrum of paradigms, ranging from regression-based techniques to diffusion-based generative approaches.

Single-Image 3D Generation Methods. We select prominent baselines representing four distinct technical trajectories within the image-to-3D domain:

- *Image-to-Gaussian:* DreamGaussian [68] and GenFusion [11].
- *Multi-view Diffusion:* SyncDreamer [9] and Ar-1-to-3 [69].
- *Video Diffusion:* SVC [70], and HunyuanWorld-Voyager [71]. Since these methods output temporal video sequences rather than explicit 3D assets, we employ a vanilla 3DGS optimization on their 80-frame generated multi-view videos to obtain the final 3D representation for a fair structural comparison.
- *Native 3D Generation:* Hunyuan3D 2.0 [53] and Trelis2 [52].

It is crucial to note that directly feeding in-the-wild, arbitrarily posed dog images into these general-purpose baselines yields heavily distorted, structurally collapsed, or semantically incorrect geometries. Therefore, to ensure the most competitive and fair comparison possible, we utilize the canonical pose image obtained by our CDOG module as the standardized input for all aforementioned single-image baselines.

Multi-Image 3D Reconstruction Methods. To further isolate and validate the specific contributions of our CA-3DGS and DCGR modules in handling dynamic/inconsistent multi-view data, we compare against robust video-driven animal reconstruction frameworks: GART [20] and BANMo [21]. For these baselines, instead of a single image, we provide the full 80-frame, 360-degree orbital video generated by our CDOG module as their input.

5.3 Comparison

Quantitative Evaluation. The quantitative comparisons against all baselines on both the Dog Synthetic Dataset and the Dog Wild Dataset are summarized in Table 1. Our proposed CORGI compellingly demonstrates state-of-the-art performance across all evaluation metrics.

On the **Dog Synthetic Dataset**, CORGI achieves the highest PSNR and second highest SSIM scores, significantly outperforming Image-to-Gaussian methods (e.g., DreamGaussian [68]) and Multi-view Diffusion models (e.g., SyncDreamer [9]). This superiority indicates that our Canonical-Driven Orbital Generation (CDOG) and Consistency-Aware Deformable 3DGS (CA-3DGS) modules successfully establish a rigorous 3D geometric structure rather than merely hallucinating independent 2D views. Nevertheless, CORGI also achieves the second lowest LPIPS score, proving its exceptional capability in preserving high-frequency textural details. Notably, Trelis2 [52] achieves the best SSIM and LPIPS scores, which we attribute to the nature of the synthetic



Fig. 4 Qualitative comparison on in-the-wild dog images. Compared to baselines that suffer from, blurry textures, or structural collapse, CORGI reconstructs high-fidelity, geometrically accurate 3D models with crisp fur details and strict multi-view consistency.

Table 1 Quantitative comparison of single-image 3D dog reconstruction. CORGI achieves the overall best performance among all baselines across both reference-based and non-reference metrics. Bold indicates the best result and underlines indicates the second-best.

Method	Dog Synthetic Dataset				Dog Wild Dataset		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP $\uparrow$	FID $\downarrow$	NIQE $\downarrow$	CLIP $\uparrow$
DreamGaussian	13.795	0.827	0.252	0.880	87.900	8.363	0.733
GenFusion	8.546	0.670	0.481	0.751	241.330	8.780	0.642
SyncDreamer	11.668	0.784	0.348	0.776	198.621	10.442	0.668
Ar-1-to-3	12.600	0.824	0.266	0.843	84.990	<u>7.421</u>	0.746
SVC	11.818	0.815	0.279	0.862	63.250	8.228	0.748
HunyuanWorld-Voyager	11.532	0.674	0.445	0.708	71.670	7.853	0.701
Hunyuan3D 2.0	12.511	0.816	0.248	<u>0.893</u>	67.850	7.922	0.749
Trellis2	13.564	0.841	0.228	0.877	<u>44.301</u>	9.040	<u>0.771</u>
GART	<u>13.799</u>	0.814	0.266	0.824	74.218	8.738	0.743
BANMo	13.749	0.811	0.262	0.854	99.654	9.409	0.718
CORGI	14.309	<u>0.836</u>	<u>0.236</u>	0.904	13.477	6.515	0.800

data: the mesh-based geometry and smooth textures of the Dog Synthetic Dataset are inherently well-suited to Trellis2’s native 3D latent representation.

On the **Dog Wild Dataset**, the non-reference metrics further highlight our method’s robustness to in-the-wild domain gaps. General-purpose generative models often struggle with the complex articulation and non-rigid deformations of real-world canines. In contrast, CORGI attains the best FID and NIQE scores, reflecting the photorealism and naturalness of our rendered views. Most importantly, our method achieves a remarkable margin in the CLIP cosine similarity score. This proves that, unlike Native 3D models that frequently collapse into generating a “generic” dog due to template bias, our framework strictly preserves the subject-specific identity, fur patterns, and structural nuances of the input image. Furthermore, compared to video-driven multi-image baselines (GART [20] and BANMo [21]), which severely degrade when optimizing directly on our pseudo-multi-view images due to generative flickering, our method robustly absorbs these inconsistencies via the view-dependent deformation field, yielding significantly higher quantitative fidelity.

Qualitative Evaluation. The visual comparisons between CORGI and representative state-of-the-art baselines are illustrated in Figure 4. The qualitative results align perfectly with our quantitative findings, explicitly revealing the inherent bottlenecks of existing paradigms when applied to highly articulated animals in the wild.

As observed in Figure 4, *Image-to-Gaussian* and *Multi-view Diffusion* methods frequently suffer from ambiguity and blurry problem. Because they lack explicit structural priors for quadrupeds, they often produce geometrically collapsed bodies, missing limbs, or floaters in occluded regions. *Video Diffusion* baselines (e.g., HunyuanWorld-Voyager [71]), while offering better temporal smoothness, lack strict 3D multi-

view constraints. When their generated 80-frame videos are lifted to vanilla 3DGS, the inherent generative inconsistencies inevitably result in torn geometries and blurry, ghosting textures.

Native 3D Generation models (e.g., Hunyuan3D 2.0 [53] and Trellis2 [52]) successfully maintain 3D consistency but exhibit a profound domain gap. Trained predominantly on synthetic or rigid game assets, they struggle to capture the complex kinematics of real dogs, often outputting rigid, toy-like geometries with completely smoothed-out fur details, failing to preserve the unique identity of the input image.

5.4 Applications

CORGI explicitly overcomes these limitations. Empowered by the D-SMAL prior within the CA-3DGS module and the progressive self-refinement of the DCGR module, our method synthesizes topologically intact, structurally coherent, and highly detailed 3D canines. Even under extreme poses, severe self-occlusions, and cluttered backgrounds, CORGI accurately reconstructs delicate anatomical structures (such as thin legs, ears, and snouts) alongside crisp, high-frequency fur textures, seamlessly bridging the gap between 2D generative priors and 3D animatable articulation.

A fundamental advantage of our generation-then-reconstruction paradigm is that it yields an inherently animatable 3D representation, directly bridging the gap between unconstrained single-image generation and downstream digital content creation. We demonstrate this practical utility through high-fidelity pose-driven animation.

Pose-Driven Animation. The final reconstructed asset from our CA-3DGS module consists of an optimized 3D Gaussian field explicitly anchored to the parametric D-SMAL template. Although our initial CDOG stage synthesizes a standardized standing pose, the individually optimized subject

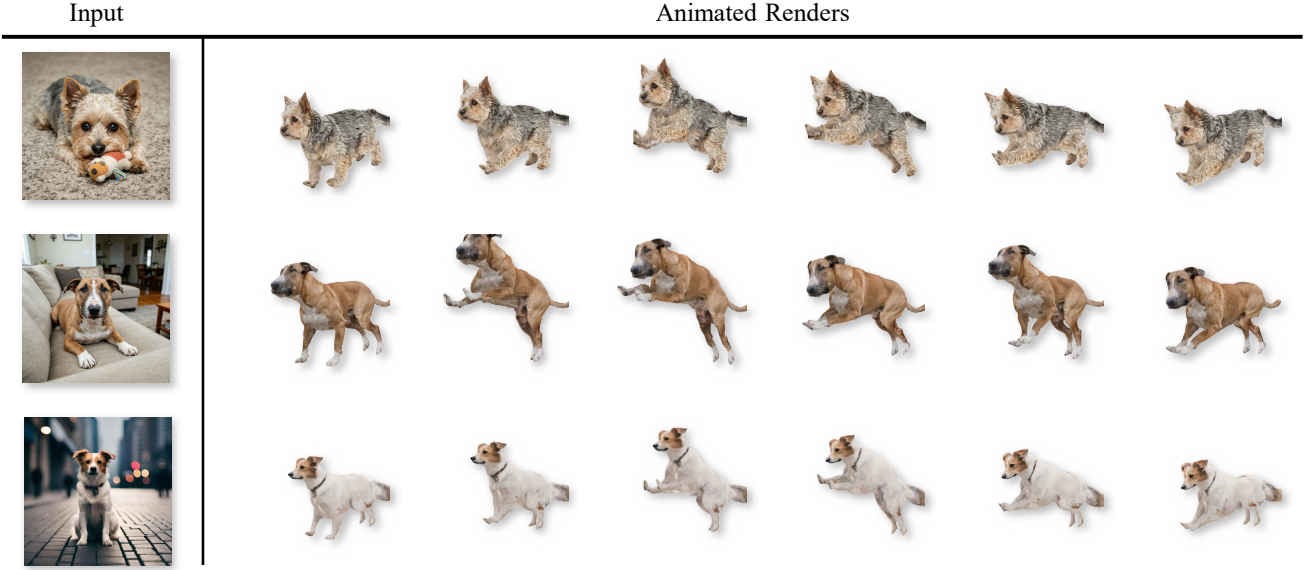


Fig. 5 Pose-driven animation results. Driven by novel D-SMAL pose sequences, our reconstructed 3D models seamlessly perform complex motions (while strictly preserving subject-specific identities, high-frequency textures, and structural integrity without geometric tearing).

may still exhibit slight kinematic deviations from a strict mathematical rest pose. To enable flawless and artifact-free animation, we first map the optimized Gaussians into a true canonical rest space via reverse Linear Blend Skinning (LBS). Specifically, for each Gaussian primitive bound to the mesh surface, we invert the local skinning transformations using the originally estimated pose parameters to extract its absolute rest-pose attributes.

Once normalized into this rigorous canonical space, the 3D dog model can be seamlessly driven by arbitrary target pose sequences. By applying forward LBS parameterized by novel D-SMAL kinematic poses, we dynamically articulate the canonical Gaussians into new poses. As illustrated in Figure 5, CORGI produces highly realistic and fluid animated sequences. Crucially, because the underlying geometry is topologically consistent and thoroughly refined by our DCGR module, the animated results exhibit zero structural tearing or floating artifacts. The model successfully preserves complex geometric nuances and high-frequency texture details (e.g., subject-specific fur patterns) even under extreme articulations and challenging novel viewpoints.

5.5 Ablation Study

To rigorously validate the efficacy of our core algorithmic designs, we conduct an incremental ablation study on the Dog Wild Dataset. We define a baseline model and systematically integrate our proposed components to analyze their distinct contributions. The quantitative results across all

configurations are summarized in Table 2.

Specifically, our evaluated configurations and their corresponding table abbreviations are defined as follows:

- **Baseline:** A vanilla 3DGS optimization applied directly to the pseudo-multi-view images generated by the CDOG module.
- **Model A:** Baseline + View-Dependent Deformation Field (Def. Field).
- **Model B:** Model A + off-the-shelf DiFix3D [56] (DiFix3D Prior).
- **Model C:** Model A + DiFix3D with LoRA fine-tuning and geometry conditioning (LoRA Fine-tune).
- **Full Model:** Model C integrated with the progressive self-refinement loop for three iterations (Refine. Loop).

Effectiveness of the Deformation Field. As shown in Figure 6, the **Baseline** produces severe artifacts and prominent structural ghosting, particularly along the body contours. This occurs because the diffusion-generated video inherently contains multi-view inconsistencies. Forcing a static, canonical Gaussian set to strictly satisfy all of these conflicting observations simultaneously leads to severe optimization conflicts. By integrating the neural deformation field, **Model A** explicitly models these cross-view discrepancies via view-specific residual adjustments, effectively disentangling the canonical geometry from generative noise. However, forcing this deformation field to generalize to a drastically different, perspective-locked view inevitably causes structural degradation, necessitating our subsequent repair modules.

Table 2 Quantitative Ablation Study. We incrementally add our proposed modules to the baseline. The Full Model achieves the best performance across all perceptual metrics.

Config	Components				Dog Wild Dataset		
	Def. Field	DiFix3D Prior	LoRA Fine-tune	Refine. Loop	FID ↓	NIQE ↓	CLIP ↑
Baseline					26.515	7.719	0.761
Model A	✓				25.934	8.201	0.781
Model B	✓	✓			22.156	7.010	0.779
Model C	✓	✓	✓		20.413	6.502	0.788
Full Model	✓	✓	✓	✓	19.703	5.754	0.795

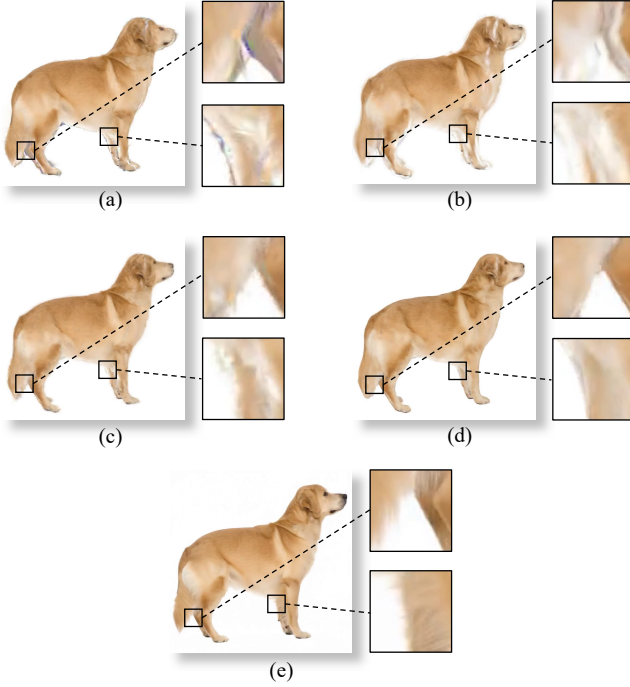


Fig. 6 Qualitative Ablation Study. (a) Baseline, (b) Model A, (c) Model B, (d) Model C, (e) Full Model

Effectiveness of LoRA Fine-Tuning. As shown in Figure 6, **Model B** attempts to repair the aforementioned structural artifacts using an off-the-shelf DiFix3D model. However, because the base DiFix3D is pre-trained exclusively on static scene datasets, it suffers from a significant domain gap. Consequently, it fails to rectify the specific degradations induced by the multi-view inconsistencies of our generated canine data. By injecting a lightweight LoRA module trained on our self-supervised dataset and incorporating explicit geometric conditioning, **Model C** successfully adapts the generic 2D diffusion prior to these targeted degradation patterns, effectively eliminating geometric tearing and texture blurring.

Effectiveness of the Progressive Self-Refinement Loop. Although Model C successfully repairs structural artifacts in single-step renderings, applying this repair independently across multi-view images cannot guarantee absolute 3D spa-

tial consistency, nor can it fully recover the high-frequency texture details of the input image. Our **Full Model** wraps this generative repair within an iterative closed-loop design. By systematically rendering, repairing, and re-optimizing the 3DGS parameters across three cycles, the self-refinement loop progressively distills authentic high-frequency details directly into the 3D geometry.

6 Conclusion

In this paper, we presented **CORGI**, a novel “generation-then-reconstruction” framework that tackles the highly challenging task of recovering high-fidelity, animatable 3D dogs from a single in-the-wild input image. To overcome the scarcity of multi-view canine data and the severe cross-view inconsistencies inherent in modern generative models, we strategically decomposed the problem into three synergistic components: governing canonical view synthesis, 3D lifting following dog prior, and geometry-guided texture refinement. Extensive evaluations demonstrate that CORGI achieves state-of-the-art reconstruction quality and generalizes seamlessly across a vast diversity of canine breeds and poses. By completely eliminating the need for paired 3D training data, our framework produces robust, artifact-free 3D models that are readily applicable to downstream animation and digital content creation.

Despite its promising results, CORGI has several limitations that point to fruitful directions for future research. First, our current animation pipeline relies on linear blend skinning (LBS) driven by the underlying D-SMAL template, which, while effective for canonical poses, struggles to produce fully realistic deformations for highly articulated or extreme postures. We believe this can be addressed by collecting and incorporating more diverse pose data to learn richer deformation priors. Second, our reconstruction framework is intrinsically tied to the D-SMAL prior, which assumes a canine-specific morphology. Extending CORGI to other quadruped species—or ideally, to arbitrary animal categories—remains an important avenue for future work, potentially by learning a more generalizable parametric animal

prior from large-scale unlabeled imagery.

References

- [1] Li Z, Amrani A, Rai S, Laga H. Advances and Trends in the 3D Reconstruction of the Shape and Motion of Animals. *arXiv preprint arXiv:2508.16062*, 2025.
- [2] Rüegg N, Tripathi S, Schindler K, Black MJ, Zuffi S. Bite: Beyond Priors for Improved Three-D Dog Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 8867–8876.
- [3] Zuffi S, Kanazawa A, Jacobs DW, Black MJ. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, 6365–6373.
- [4] Sabathier R, Mitra NJ, Novotny D. Animal Avatars: Reconstructing Animatable 3D Animals from Casual Videos. In *European Conference on Computer Vision*, 2024, 270–287.
- [5] Li Z, Litvak D, Li R, Zhang Y, Jakab T, Rupprecht C, Wu S, Vedaldi A, Wu J. Learning the 3d Fauna of the Web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 9752–9762.
- [6] Yao CH, Hung WC, Li Y, Rubinstein M, Yang MH, Jampani V. Lassie: Learning Articulated Shapes from Sparse Image Ensemble via 3D Part Discovery. *Advances in Neural Information Processing Systems*, 2022, 35: 15296–15308.
- [7] Wu S, Li R, Jakab T, Rupprecht C, Vedaldi A. Magicpony: Learning Articulated 3D Animals in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 8792–8802.
- [8] Liu R, Wu R, Van Hoorick B, Tokmakov P, Zakharov S, Vondrick C. Zero-1-to-3: Zero-shot one Image to 3D Object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, 9298–9309.
- [9] Liu Y, Lin C, Zeng Z, Long X, Liu L, Komura T, Wang W. Syncdreamer: Generating Multiview-Consistent Images from a Single-View Image. In *International Conference on Learning Representations*, volume 2024, 2024, 27676–27697.
- [10] Liu M, Shi R, Chen L, Zhang Z, Xu C, Wei X, Chen H, Zeng C, Gu J, Su H. One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion. *arXiv preprint arXiv:2311.07885*, 2023.
- [11] Wu S, Xu C, Huang B, Geiger A, Chen A. Genfusion: Closing the Loop between Reconstruction and Generation via Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, 6078–6088.
- [12] Ren X, Shen T, Huang J, Ling H, Lu Y, Nimier-David M, Müller T, Keller A, Fidler S, Gao J. Gen3c: 3D-Informed World-Consistent Video Generation with Precise Camera Control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, 6121–6132.
- [13] Cho G, Kang C, Soon D, Joo K. DogRecon: Canine Prior-Guided Animatable 3D Gaussian Dog Reconstruction From A Single Image. *International Journal of Computer Vision*, 2025, 133(9): 6332–6346.
- [14] Kerbl B, Kopanas G, Leimkühler T, Drettakis G, et al.. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics (TOG)*, 2023, 42(4): 139–1.
- [15] Li C, Ghorbani N, Broomé S, Rashid M, Black MJ, Hernlund E, Kjellström H, Zuffi S. Hsmal: Detailed Horse Shape and Pose Reconstruction for Motion Pattern Recognition. *arXiv preprint arXiv:2106.10102*, 2021.
- [16] Zuffi S, Mellbin Y, Li C, Hoeschle M, Kjellström H, Polikovskiy S, Hernlund E, Black MJ. Varen: Very Accurate and Realistic Equine Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 5374–5383.
- [17] Yang G, Sun D, Jampani V, Vlasic D, Cole F, Chang H, Ramanan D, Freeman WT, Liu C. Lasr: Learning Articulated Shape Reconstruction from a Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 15980–15989.
- [18] Rueegg N, Zuffi S, Schindler K, Black MJ. Barc: Learning to Regress 3D Dog Shape from Images by Exploiting Breed Information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 3876–3884.
- [19] Lyu J, Zhu T, Gu Y, Lin L, Cheng P, Liu Y, Tang X, An L. Animer: Animal Pose and shape Estimation using Family Aware Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, 17486–17496.
- [20] Lei J, Wang Y, Pavlakos G, Liu L, Daniilidis K. Gart: Gaussian Articulated Template Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 19876–19887.
- [21] Yang G, Vo M, Neverova N, Ramanan D, Vedaldi A, Joo H. Banmo: Building Animatable 3D Neural Models from Many Casual Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 2863–2873.
- [22] Yao CH, Hung WC, Li Y, Rubinstein M, Yang MH, Jampani V. Hi-Lassie: High-Fidelity Articulated Shape and Skeleton Discovery from Sparse Image Ensemble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 4853–4862.
- [23] Liu D, Stathopoulos A, Zhangli Q, Gao Y, Metaxas D. Leopard: Learning Explicit Part Discovery for 3D Articulated Shape Reconstruction. *Advances in Neural Information Processing Systems*, 2023, 36: 54187–54198.
- [24] Yao CH, Raj A, Hung WC, Rubinstein M, Li Y, Yang MH, Jampani V. Artic3d: Learning Robust Articulated 3D Ahapes from Noisy Web Image Collections. *Advances in Neural Information Processing Systems*, 2023, 36: 48173–48184.
- [25] Wu Y, Chen Z, Liu S, Ren Z, Wang S. Casa: Category-Agnostic Skeletal Animal Reconstruction. *Advances in Neural Information Processing Systems*, 2022, 35: 28559–28574.
- [26] Kaye B, Jakab T, Wu S, Rupprecht C, Vedaldi A. Dualpm:



- Dual Posed-Canonical Point Maps for 3D Shape and Pose Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, 6425–6435.
- [27] Wang C, Peng HY, Liu YT, Gu J, Hu SM. Diffusion Models for 3D Generation: A Survey. *Computational Visual Media*, 2025, 11(1): 1–28.
- [28] Tatarchenko M, Richter SR, Ranftl R, Li Z, Koltun V, Brox T. What do Single-View 3D Reconstruction Networks Learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 3405–3414.
- [29] Fu K, Peng J, He Q, Zhang H. Single Image 3D Object Reconstruction Based on Deep Learning: A Review. *Multimedia Tools and Applications*, 2021, 80: 463–498.
- [30] Kato H, Harada T. Learning view Priors for Single-View 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, 9778–9787.
- [31] Li X, Liu S, Kim K, De Mello S, Jampani V, Yang MH, Kautz J. Self-Supervised Single-View 3D Reconstruction via Semantic Consistency. In *European Conference on Computer Vision*, 2020, 677–693.
- [32] Fahim G, Amin K, Zarif S. Single-View 3D reconstruction: A Survey of Deep Learning Methods. *Computers & Graphics*, 2021, 94: 164–190.
- [33] Schwarz K, Liao Y, Niemeyer M, Geiger A. Graf: Generative Radiance Fields for 3D-Aware Image Synthesis. *Advances in Neural Information Processing Systems*, 2020, 33: 20154–20166.
- [34] Niemeyer M, Geiger A. Giraffe: Representing Scenes as Compositional Generative Neural Feature Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, 11453–11464.
- [35] Gu J, Liu L, Wang P, Theobalt C. Stylenerf: A Style-Based 3D-Aware Generator for High-Resolution Image Synthesis. *arXiv preprint arXiv:2110.08985*, 2021.
- [36] Chan ER, Lin CZ, Chan MA, Nagano K, Pan B, De Mello S, Gallo O, Guibas LJ, Tremblay J, Khamis S, et al.. Efficient Geometry-Aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 16123–16133.
- [37] Poole B, Jain A, Barron JT, Mildenhall B. Dreamfusion: Text-to-3D using 2D Diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
- [38] Liang Y, Yang X, Lin J, Li H, Xu X, Chen Y. Luciddreamer: Towards High-Fidelity Text-to-3D Generation via Interval Score Matching. *arXiv preprint arXiv:2311.11284*, 2023.
- [39] Qian G, Mai J, Hamdi A, Ren J, Siarohin A, Li B, Lee HY, Skorokhodov I, Wonka P, Tulyakov S, et al.. Magic123: One Image to High-Quality 3D Object Generation using both 2D and 3D Diffusion Priors. In *International Conference on Learning Representations*, volume 2024, 2024, 48142–48159.
- [40] Wang Z, Lu C, Wang Y, Bao F, Li C, Su H, Zhu J. Prolificdreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation. *Advances in Neural Information Processing Systems*, 2024, 36.
- [41] Shi Y, Wang P, Ye J, Mai L, Li K, Yang X. MVDream: Multi-View Diffusion for 3D Generation. In *International Conference on Learning Representations*, volume 2024, 2024, 39838–39859.
- [42] Long X, Guo YC, Lin C, Liu Y, Dou Z, Liu L, Ma Y, Zhang SH, Habermann M, Theobalt C, et al.. Wonder3d: Single Image to 3D using Cross-Domain Diffusion. *CVPR*, 2024.
- [43] Hong Y, Zhang K, Gu J, Bi S, Zhou Y, Liu D, Liu F, Sunkavalli K, Bui T, Tan H. Lrm: Large Reconstruction Model for Single Image to 3D. *arXiv preprint arXiv:2311.04400*, 2023.
- [44] Wang P, Tan H, Bi S, Xu Y, Luan F, Sunkavalli K, Wang W, Xu Z, Zhang K. Pf-Lrm: Pose-Free Large Reconstruction Model for Joint Pose and Shape Prediction. *arXiv preprint arXiv:2311.12024*, 2023.
- [45] Wang Z, Wang Y, Chen Y, Xiang C, Chen S, Yu D, Li C, Su H, Zhu J. Crm: Single Image to 3D Textured Mesh with Convolutional Reconstruction Model. *arXiv preprint arXiv:2403.05034*, 2024.
- [46] Tang J, Chen Z, Chen X, Wang T, Zeng G, Liu Z. LGM: Large Multi-View Gaussian Model for High-Resolution 3D Content Creation. *arXiv preprint arXiv:2402.05054*, 2024.
- [47] Xu J, Cheng W, Gao Y, Wang X, Gao S, Shan Y. InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-View Large Reconstruction Models. *arXiv preprint arXiv:2404.07191*, 2024.
- [48] Tochilkin D, Pankratz D, Liu Z, Huang Z, Letts A, Li Y, Liang D, Laforte C, Jampani V, Cao YP. Tripotr: Fast 3D Object Reconstruction from a Single Image. *arXiv preprint arXiv:2403.02151*, 2024.
- [49] Liu M, Xu C, Jin H, Chen L, Varma T M, Xu Z, Su H. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. *Advances in Neural Information Processing Systems*, 2024, 36.
- [50] Zhang L, Wang Z, Zhang Q, Qiu Q, Pang A, Jiang H, Yang W, Xu L, Yu J. Clay: A Controllable Large-Scale Generative Model for Creating High-Quality 3D Assets. *ACM Transactions on Graphics (TOG)*, 2024, 43(4): 1–20.
- [51] Xiang J, Lv Z, Xu S, Deng Y, Wang R, Zhang B, Chen D, Tong X, Yang J. Structured 3D Latents for Scalable and Versatile 3D Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, 21469–21480.
- [52] Xiang J, Chen X, Xu S, Wang R, Lv Z, Deng Y, Zhu H, Dong Y, Zhao H, Yuan NJ, et al.. Native and Compact Structured Latents for 3D Generation. *arXiv preprint arXiv:2512.14692*, 2025.
- [53] Zhao Z, Lai Z, Lin Q, Zhao Y, Liu H, Yang S, Feng Y, Yang M, Zhang S, Yang X, et al.. Hunyuan3d 2.0: Scaling Diffusion



- Models for High Resolution Textured 3d Assets Generation. *arXiv preprint arXiv:2501.12202*, 2025.
- [54] Zwicker M, Pfister H, Van Baar J, Gross M. Ewa Volume Splatting. In *Proceedings Visualization, 2001. VIS'01.*, 2001, 29–538.
- [55] Wei J, Leutenegger S, Schaefer S. GSFix3D: Diffusion-Guided Repair of Novel Views in Gaussian Splatting. *arXiv preprint arXiv:2508.14717*, 2025.
- [56] Wu JZ, Zhang Y, Turki H, Ren X, Gao J, Shou MZ, Fidler S, Gojcic Z, Ling H. Difix3d+: Improving 3D Reconstructions with Single-Step Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, 26024–26035.
- [57] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W. LoRA: Low-Rank Adaptation of Large Language Models, 2021.
- [58] Wu C, Li J, Zhou J, Lin J, Gao K, Yan K, Yin Sm, Bai S, Xu X, Chen Y, et al.. Qwen-Image Technical Report. *arXiv preprint arXiv:2508.02324*, 2025.
- [59] Wan T, Wang A, Ai B, Wen B, Mao C, Xie CW, Chen D, Yu F, Zhao H, Yang J, Zeng J, Wang J, Zhang J, Zhou J, Wang J, Chen J, Zhu K, Zhao K, Yan K, Huang L, Feng M, Zhang N, Li P, Wu P, Chu R, Feng R, Zhang S, Sun S, Fang T, Wang T, Gui T, Weng T, Shen T, Lin W, Wang W, Wang W, Zhou W, Wang W, Shen W, Yu W, Shi X, Huang X, Xu X, Kou Y, Lv Y, Li Y, Liu Y, Wang Y, Zhang Y, Huang Y, Li Y, Wu Y, Liu Y, Pan Y, Zheng Y, Hong Y, Shi Y, Feng Y, Jiang Z, Han Z, Wu ZF, Liu Z. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.
- [60] Wang Z, Chen S, Mo L, Gao X, Shen Y, Ding L, Liang W. DogMo: A Large-Scale Multi-View RGB-D Dataset for 4D Canine Motion Recovery. *arXiv preprint arXiv:2510.24117*, 2025.
- [61] Liu S, Li T, Chen W, Li H. Soft Rasterizer: A Differentiable Renderer for Image-based 3D Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 7708–7717.
- [62] Jiang Y, Liao Q, Li X, Ma L, Zhang Q, Zhang C, Lu Z, Shan Y. Uv Gaussians: Joint Learning of Mesh Deformation and Gaussian Textures for Human Avatar Modeling. *Knowledge-Based Systems*, 2025, 320: 113470.
- [63] Gao L, Yang J, Zhang BT, Sun JM, Yuan YJ, Fu H, Lai YK. Real-Time Large-Scale Deformation of Gaussian Splatting. *ACM Transactions on Graphics (TOG)*, 2024, 43(6): 1–17.
- [64] Igarashi T, Moscovich T, Hughes JF. As-Rigid-as-Possible Shape Manipulation. *ACM Transactions on Graphics (TOG)*, 2005, 24(3): 1134–1141.
- [65] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, 586–595.
- [66] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems*, 2017, 30: 1–12.
- [67] Zhang L, Zhang L, Bovik AC. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Transactions on Image Processing*, 2015, 24(8): 2579–2591.
- [68] Tang J, Ren J, Zhou H, Liu Z, Zeng G. Dreamgaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. In *International Conference on Learning Representations*, volume 2024, 2024, 33879–33896.
- [69] Zhang X, Zhou Y, Wang K, Wang Y, Li Z, Jiao S, Zhou D, Hou Q, Cheng MM. Ar-1-to-3: Single Image to Consistent 3D Object via Next-View Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, 26273–26283.
- [70] Zhou J, Gao H, Voleti V, Vasishtha A, Yao CH, Boss M, Torr P, Rupprecht C, Jampani V. Stable Virtual Camera: Generative View Synthesis with Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, 12405–12414.
- [71] Huang T, Zheng W, Wang T, Liu Y, Wang Z, Wu J, Jiang J, Li H, Lau RW, Zuo W, Guo C. Voyager: Long-Range and World-Consistent Video Diffusion for Explorable 3D Scene Generation. *arXiv preprint arXiv:2506.04225*, 2025.

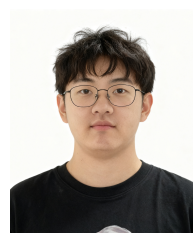
Author biography



Yuxiao Wu is a Ph.D. student in School of Mathematical Sciences, University of Science and Technology of China. His current research focuses on 3D animal reconstruction, single-image reconstruction and virtual reality.



Boyi Zhu is an undergraduate student at the School of Mathematical Sciences, University of Science and Technology of China. His current research interests include computer graphics and 3D reconstruction.



Weile Li is an undergraduate student at the School of Artificial Intelligence and Data Science, University of Science and Technology of China. His current research interests include computer graphics, multimodal large models, and etc.



computer graphics and embodied intelligence.

Yumeng Liu is a Postdoctoral Researcher at the School of Mathematical Sciences, University of Science and Technology of China. She received her Ph.D. in Computer Science from the University of Hong Kong. Her research interests include 3D/4D reconstruction, com-



a postdoctor at University of Science and Technology of China. His research interests include 3D reconstruction, computer vision, and machine learning.

Youcheng Cai YOUCHENG CAI received the B.S. degree in information and computing science and Ph.D. degree in computer application technology from the Hefei University of Technology, Hefei, China, in 2008 and 2023, respectively. He is currently working as



His research interests include geometric processing and computer-aided geometric design. His research work can be found at his website: <http://staff.ustc.edu.cn/~fuxm/geometric design>.

XiaoMing Fu received a BSc degree in 2011 and a Ph.D. in 2016 from the University of Science and Technology of China. He is an associate professor at the School of Mathematical Sciences, University of Science and Technology of China. His research



He paid academic visits to Harvard University in 2009 and 2011. His research interests include geometric processing and image processing.

Ligang Liu received his B.Sc. and Ph.D. degrees from Zhejiang University, China, in 1996 and 2001, respectively. He is a professor at USTC. Between 2001 and 2004, he worked at Microsoft Research Asia, and he worked at Zhejiang University from 2004 to 2012.